

A UNIFIED STATISTICAL FRAMEWORK FOR RNA SEQUENCE DATA FROM INDIVIDUAL CELLS AND TISSUE

BY LINGXUE ZHU*, JING LEI* AND KATHRYN ROEDER*

*Carnegie Mellon University**

Recent advances in technology have enabled the measurement of RNA levels for individual cells. Compared to traditional tissue-level RNA-seq data, single cell sequencing yields valuable insights about gene expression profiles for different cell types, which is potentially critical for understanding many complex human diseases. However, developing quantitative tools for such data remains challenging because of high levels of technical noise, especially the “dropout” events. A “dropout” happens when the RNA for a gene fails to be amplified prior to sequencing, producing a “false” zero in the observed data. In this paper, we propose a unified statistical framework for both single cell and tissue RNA-seq data, formulated as a hierarchical model. Our framework borrows the strength from both data sources and carefully models the dropouts in single cell data, leading to a more accurate estimation of cell type specific gene expression profile. In addition, our model naturally provides inference on (i) the dropout entries in single cell data that need to be imputed for downstream analyses; and (ii) the mixing proportions of different cell types in tissue samples. We adopt an empirical Bayes approach, where parameters are estimated using the EM algorithm and approximate inference is obtained by Gibbs sampling. Simulation results illustrate that our framework outperforms existing approaches both in correcting for dropouts in single cell data, as well as in deconvolving tissue samples. We also demonstrate an application to gene expression data on fetal brains, where our model successfully imputes the dropout genes and reveals cell type specific expression patterns.

1. Introduction.

1.1. *Single cell RNA-seq data.* RNA sequencing (RNA-seq) is a powerful tool to measure gene expression levels throughout the genome. With sufficient sequencing depth, even weakly expressed transcripts can be accurately captured. Traditional RNA-seq experiments are conducted on tissues that include a population of cells; however, the different gene expression profiles for distinct cell types can have profound functional consequences. Therefore, characterizing cell type specific gene expression patterns can be critical for understanding the development of tissues as well as revealing molecular mechanisms for many human diseases. Recent advances in high-throughput technologies have made it possible to profile hundreds and thousands of individual cells, yielding a high-resolution view of previously uncharacterized cellular states within a population (Kolodziejczyk et al., 2015; Fan et al., 2015). With single cell RNA-seq data, one can investigate distinct subpopulations of cells, gain better understanding of the developmental features of different

Keywords and phrases: Single cell RNA sequencing, hierarchical model, empirical Bayes, Gibbs sampling, EM algorithm.

cell types (Grün et al., 2015), identify cellular differences between healthy and diseased tissues (Kharchenko et al., 2014), and infer gene-regulatory interactions (Padovan-Merhar & Raj, 2013).

The challenges of modeling single cell RNA-seq data come from high cell-to-cell variation, as well as high levels of technical noise during sequencing due to the low amounts of starting mRNAs in individual cells. One important bias comes from the so-called “dropout” events. A dropout happens when a transcript is not detected due to failure of amplification prior to sequencing, leading to a “false” zero in the observed data (Kolodziejczyk et al., 2015). Given the excessive amount of zero observations in single cell RNA-seq data, it is critical to distinguish between (i) the dropout genes where transcripts are missed in sequencing; and (ii) the “structural zeros” where the genes are truly un-expressed. Modeling the dropout events is especially challenging because of their complicated dependency on gene expression levels and cell characteristics. Specifically, dropouts are more likely to occur in genes expressed at low levels, and certain cells may have systematically higher dropout probabilities than others. In addition to dropout events, other challenges in modeling single cell data include the over-dispersion due to both cellular and technical variation, as well as high magnitude outliers due to bursts and fluctuations of gene expression levels.

Despite the success of many early single-cell studies, statistical tools that account for the technical noise in single cell RNA-seq data, especially the dropout events, are limited. There have been efforts to analyze single cell data for various purposes. Many methods propose to quantify and account for technical noise using spike-ins (Brennecke et al., 2013; Vallejos et al., 2015, 2016). However, spike-ins are rarely available in single cell data due to its expenses in practice. For differential expression analysis, Kharchenko et al. (2014) propose SCDE, a Bayesian hypothesis testing procedure using a three-component mixture model to capture technical noise; subsequently, Finak et al. (2015) propose MAST, using a hurdle model that can adjust for various covariates; more recently, Vu et al. (2016) construct a beta-poisson mixture model, integrated within a generalized linear model framework. Various relevant models for these data have been proposed. Satija et al. (2015) consider the problem of inferring the spatial localization of single cells in complex tissues; Pierson & Yau (2015) propose Zero-Inflated Factor Analysis (ZIFA) for dimension reduction, and Prabhakaran et al. (2016) focus on clustering unlabeled single cells while accounting for technical variation. All of these aforementioned methods have been successfully applied to different single cell data sets. However, analytical methods that aim at the fundamental problem of imputing dropout genes and estimating the cell-type-specific gene expression profiles remain underdeveloped. In addition, the possibility of utilizing the information in tissue RNA-seq data is unexplored.

1.2. Tissue RNA-seq data. Compared to single cell data, traditional tissue-level RNA-seq data provides more reliable measurements of gene expression by having deeper sequencing depths. Thus it does not suffer from dropout events. To characterize cell type specific gene expression profiles from the tissue-level measurements, however, an extra step of deconvolution is required. Because the mixing proportions of different cell types in each tissue sample are usually unknown, deconvolution is

a nontrivial problem.

One way to approach the deconvolution problem is to start from nonnegative matrix factorization (NMF). Consider an observed gene expression matrix $X \in \mathbb{R}^{N \times M}$ for N genes in M tissue samples, each containing K different cell types. The goal is to find two non-negative matrices $\tilde{A} \in \mathbb{R}^{N \times K}$ and $W \in \mathbb{R}^{K \times M}$, such that

$$X \approx \tilde{A}W, \quad (1.1)$$

where each column of W sums to one, which represents the mixing proportion of K cell types in each tissue sample, and the k -th column of \tilde{A} represents the average gene expression levels for the k -th type of cells.

A fundamental challenge of solving problem (1.1) is that, in general, the factorization cannot guarantee to be unique. To obtain a biologically meaningful result, a set of “marker genes” is needed to guide the algorithm. A marker gene is a gene that is only expressed in one of the K cell types. In other words, there are several rows of \tilde{A} that are priorly known to be non-zero only in one column. This is equivalent to the separability assumption introduced by Donoho & Stodden (2003) for uniqueness of NMF.

If a set of marker genes is known for each cell type, solving the deconvolution problem (1.1) is straightforward. However, this is rarely the case in practice. In fact, extracting high-quality marker genes is a challenging step, which is often approached by analyzing purified cells to identify a set of genes that have significantly different expression levels across different cell types (Abbas et al., 2009). However, the estimated marker genes are usually noisy, in the sense that the corresponding rows of \tilde{A} may contain many small numbers instead of zeros. Unfortunately, most existing deconvolution methods tend to be sensitive to such noise, including the Deconf algorithm (Repsilber et al., 2010), semi-supervised Nonnegative Matrix Factorization algorithm (ssNMF) (Gaujoux & Seoighe, 2012), and Digital Sorting Algorithm (DSA) (Zhong et al., 2013). With the emergence of single cell data, it is now possible to develop a new deconvolution algorithm that avoids the problem of estimating marker genes by utilizing the cellular expression profiles in an integrated way.

1.3. Our contribution. In this paper, we propose a unified statistical framework for both single cell and tissue RNA-seq data. Borrowing the strength from both data sources, our goal is to (i) obtain reliable estimation of cell type specific gene expression profiles; (ii) infer the dropout entries in single cell data; and (iii) infer the mixing proportions of different cell types in tissue samples. Our framework explicitly models the dropout events in single cell data, and captures the relationship between dropout probability and the true gene expression levels. By involving high-quality tissue data, our model achieves more accurate estimation of cellular expression profiles than using only single cell data. By incorporating the single cell data, our framework provides, for the first time, deconvolution of the tissue samples without going through the error-prone procedure of estimating marker genes. To the best of our knowledge, this is the first model that jointly analyzes these two types of RNA-seq data. We will illustrate in simulation (Section 4) and real-world data

(Section 5) that our model successfully corrects for the dropouts in single cell data, and provides reliable deconvolution for tissue samples.

2. A Unified Statistical Model. Suppose RNA-seq is conducted on N genes and K types of cells are of interest. Then tissue and single cell RNA-seq data can be linked together by a common profile matrix $A \in \mathbb{R}^{N \times K}$, where the k -th column $A_{\cdot k}$ represents the expected *relative* expression levels of N genes in the k -th type of cells, such that each column sums to one. Note that by considering the *relative* expression levels, the profile matrix A does not depend on sequencing depths, and thus remains the same in both data sources. The two data sources provide two different views on the profile matrix A . In single cell data, the observations are independent realizations of different columns of A with extra noise due to dropout events. In tissue data, the expected relative expression levels for a mixture sample are weighted sums of columns of A , where the weights correspond to mixing proportions of different cell types. Here, we propose a unified framework to analyze the tissue and single cell RNA-seq data together, which borrows the strength from both data sets and achieves more accurate estimation on the profile matrix. This further enhances the performance of deconvolving tissue samples, as well as inferring and imputing the dropout genes in single cells.

The plate model for generating single cell and tissue RNA-seq data is given in Figure 1. Specifically, for single cell data, let $Y \in \mathbb{R}^{N \times L}$ represent the measured expression levels of N genes in L single cells, where the entries are RNA-seq counts. To model the dropout events, we introduce the binary observability variable $S \in \{0, 1\}^{N \times L}$, where $S_{il} = 0$ if gene i in cell l is dropped out, and $S_{il} = 1$ if it is properly amplified. For each cell l , let $G_l \in \{1, \dots, K\}$ denote its type, then the vector of gene expression $Y_{\cdot l} \in \mathbb{R}^N$ is assumed to follow a Multinomial distribution with probability vector p_l . Without dropout events, p_l would be the corresponding column of the profile matrix, $A_{\cdot G_l}$, which is the true relative expression levels for cell type G_l . With the existence of dropouts, p_l becomes the element-wise product of $A_{\cdot G_l}$ and $S_{\cdot l}$, which is then normalized to sum to one. To capture the dependency between dropout probabilities and gene expression levels, the observation probability $\pi_{il} = \mathbb{P}(S_{il} = 1)$ is modeled as a logistic function of a linear transformation of A_{i, G_l} , with the intercept and slope being (κ_l, τ_l) , which are cell-dependent to reflect cellular heterogeneity. Under this model, the set of dropout entries and structural zeros can be written as

$$\begin{aligned} \text{dropouts} &= \{(i, l) : S_{il} = 0\}, \text{ and} \\ \text{structural zeros} &= \{(i, l) : S_{il} = 1, Y_{il} = 0\}. \end{aligned} \tag{2.1}$$

For tissue data, let $X \in \mathbb{R}^{N \times M}$ represent the RNA-seq counts of N genes in M tissue samples. For the j -th tissue sample, let $W_{\cdot j} \in \mathbb{R}^K$ denote the mixing proportions of K cell types in the sample, satisfying $\sum_{k=1}^K W_{kj} = 1$. Then the gene expression vector $X_{\cdot j} \in \mathbb{R}^N$ is assumed to also follow a Multinomial distribution, where the probability vector is the weighted sum of K columns of A with the weights being $W_{\cdot j}$.

For the hierarchical model setting, we assign the conjugate Dirichlet prior for the

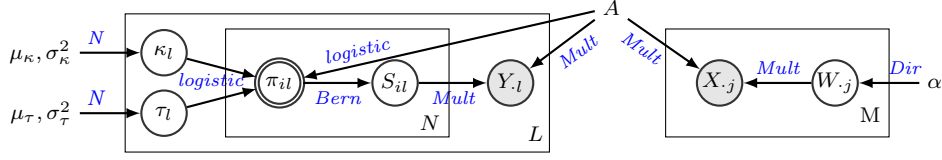


FIG. 1. Plate model for a unified framework of single cell data (on the left) and tissue samples (on the right). The two greyed nodes X and Y represent observed gene expression levels. Node S is a binary variable representing dropout status in single cells, and node W represents the mixing proportions in tissue samples. The node π representing observation probability is double-circled since it is deterministic, and all model parameters are shown without circles, including the profile matrix A that links the two data sources.

mixing proportions $W_{\cdot j}$, and Gaussian priors for the cell-dependent dropout parameters (κ_l, τ_l) . Here, we adopt an empirical Bayes approach, where the parameters are estimated by maximum-likelihood-estimations (MLE) using an expectation-maximization (EM) algorithm. Using this framework, our goals can be specified as: (i) learn the profile matrix A as part of the model parameters, which characterizes the cellular gene expression profiles; (ii) make posterior inference on the dropout status S for single cell data, which can be used to infer dropout entries, and (iii) make posterior inference on the mixing proportions W in tissue samples. Finally, the inferred dropout entries in single cell data can be imputed by their expected values using the estimated A and sequencing depths R_l .

Full model specification..

- Tissue data
 - $W_{\cdot j} \stackrel{i.i.d.}{\sim} \text{Dirichlet}(\alpha)$ for $j = 1, \dots, M$, where $\alpha \in \mathbb{R}^K$, $\alpha \geq 0$.
 - $X_{\cdot j} | W_{\cdot j} \stackrel{indep.}{\sim} \text{Multinomial}(R_j, AW_{\cdot j})$ for $j = 1, \dots, M$, where $R_j = \sum_{i=1}^N X_{ij}$.
- Single cell data
 - $\kappa_l \stackrel{i.i.d.}{\sim} N(\mu_\kappa, \sigma_\kappa^2)$, $\tau_l \stackrel{i.i.d.}{\sim} N(\mu_\tau, \sigma_\tau^2)$ for $l = 1, \dots, L$.
 - $\pi_{il} = \text{logistic}(\kappa_l + \tau_l A_{i, G_l})$, where $G_l \in \{1, \dots, K\}$ is the type of the l -th cell.
 - $S_{il} | \kappa_l, \tau_l \stackrel{indep.}{\sim} \text{Bernoulli}(\pi_{il})$ for $i = 1, \dots, N$; $l = 1, \dots, L$.
 - $Y_{\cdot l} | S_{\cdot l} \stackrel{indep.}{\sim} \text{Multinomial}(R_l, p_l)$ for $l = 1, \dots, L$, where $R_l = \sum_{i=1}^N Y_{il}$,

$$p_l = (p_{il})_{i=1, \dots, N}, \text{ where } p_{il} = \frac{A_{i, G_l} S_{il}}{\sum_{n=1}^N A_{n, G_l} S_{nl}}.$$

Remark 1. We assume all entries in A to be strictly positive. In principle, one can allow some entries A_{ik} to be exactly zero, but this will lead to a degenerate multinomial distribution and complicate the likelihood function. In addition, making inference on S_{il} when $A_{i, G_l} = 0$ is an ill-defined problem. Finally, if $A_{ik} = 0$,

then we will have $X_{il} = 0$ for all type- k cells. However, such clean structures rarely appear in real data. In practice, it is usually helpful to use small positive numbers rather than exact zeros to capture the background signal in sequencing processes (Kharchenko et al., 2014).

Remark 2. Although we only present the full model for unified analysis here, it is straightforward to use one part of the model when only one data source is available. Note that when only tissue data are available, extra information about marker genes needs to be incorporated in the model to avoid the non-identifiability issue as mentioned in Section 1.2.

3. Inference and Estimation: EM Algorithm. This section presents an expectation-maximization (EM) algorithm (Dempster et al., 1977) for fitting the maximum likelihood estimation (MLE) of the parameters $\theta = (A, \alpha, \mu_\kappa, \sigma_\kappa^2, \mu_\tau, \sigma_\tau^2)$, as well as a Gibbs sampling algorithm for posterior inference on latent variables $H = (W, S, \kappa, \tau)$. As illustrated in Section 2, the key values of scientific interests include (i) an estimate of the profile matrix A that characterizes the cellular gene expression profiles; (ii) $\mathbb{E}[S|Y, \theta]$, the inferred dropout probability at each entry in single cell data; and (iii) $\mathbb{E}[W|X, \theta]$, the inferred mixture proportion of tissue samples.

The main difficulty of handling our model is the intractable posterior distributions due to non-conjugacy. Therefore, approximate inference needs to be performed. One of the main methods for approximate inference in Bayesian modeling is Monte Carlo Markov Chain (MCMC) sampling (Gelfand & Smith, 1990), where a Markov chain on latent variables is constructed, with stationary distribution being the true posterior. After obtaining a long enough chain, the posterior can be approximated with empirical estimation. Gibbs sampling (Geman & Geman, 1984; Casella & George, 1992) is one of the most widely used forms of MCMC algorithms given its simplicity and efficiency. On the other hand, variational methods form an alternative line for approximate inference, where the posterior is approximated analytically by a family of tractable distributions (Jordan et al., 1999; Wainwright & Jordan, 2008; Blei et al., 2016). While being computationally scalable in many large-scale problems, variational methods are inherently less accurate due to the inevitable gap between the variational distributions and the true posterior distribution.

In this paper, we present a Gibbs sampling algorithm for approximate inference on latent variables using the data augmentation trick. This algorithm can also be used in the E-step of the EM procedure, leading to a Gibbs-EM (GEM) algorithm for obtaining MLEs of model parameters (?). The specific steps are outlined in Section 3.1 and Section 3.2, and more details can be found in the appendix. Finally, we point out that one can also proceed with variational inference, but due to space limitation, we do not pursue this approach in detail.

3.1. E-step: Gibbs sampling. The latent variables for tissue data and single cell data are conditionally independent given observed data X, Y and parameters. Therefore, Gibbs sampling can be performed on the two data sources in parallel. In this section, we describe the sampling procedure for the two parts separately.

Tissue data. To obtain the posterior inference of W (the mixing proportions) in tissue data, we re-write the model to be mixture of multinomials by introducing the augmented latent variables Z and d as follows:

$$\begin{aligned} W_{.j} &\stackrel{i.i.d.}{\sim} \text{Dirichlet}(\alpha), \quad j = 1, \dots, M, \\ Z_{rj} &\stackrel{i.i.d.}{\sim} \text{Multinomial}(1, W_{.j}), \quad r = 1, \dots, R_j, \\ d_{rj} &\stackrel{indep.}{\sim} \text{Multinomial}(1, A_{.Z_{rj}}), \quad r = 1, \dots, R_j, \\ X_{ij} &= \sum_{r=1}^{R_j} I_{\{d_{rj}=i\}}, \quad i = 1, \dots, N, \quad j = 1, \dots, M. \end{aligned} \quad (3.1)$$

Note that this model is closely related to the Latent Dirichlet Allocation (LDA) model (Blei et al., 2003) in topic modeling, if we view a gene as a word, a cell type as a topic, and a tissue sample as a document. Although the Gibbs sampling algorithm has been developed for LDA in Griffiths & Steyvers (2004), there are two difficulties that prevent us from directly applying this algorithm to our model. First, the LDA model assumes observations of d_{rj} , which are the actual words in an document, but in RNA-seq data, only the final counts X_{ij} are observed. Second, the sequencing depths R_j 's are typically large in real data, so it will be extremely computationally demanding to keep track of Z_{rj} and d_{rj} . Therefore, we propose a modified algorithm by defining another set of augmented latent variables

$$\tilde{Z}_{ij,k} := \sum_{r:d_{rj}=i} I_{\{Z_{rj}=k\}} \text{ and } \tilde{Z}_{ij} := (\tilde{Z}_{ij,k}) \in \mathbb{R}^K, \quad (3.2)$$

and it can be shown that

$$\begin{aligned} W_{.j} | W_{.(-j)}, \tilde{Z}, X &\sim \text{Dirichlet} \left(\alpha + \sum_{i=1}^N \tilde{Z}_{ij} \right), \\ \tilde{Z}_{ij} | \tilde{Z}_{(-ij)}, W, X &\sim \text{Multinomial} \left(X_{ij}, \frac{A_{i.} \odot W_{.j}}{\sum_{k=1}^K A_{ik} W_{kj}} \right), \end{aligned} \quad (3.3)$$

where \odot denotes element-wise multiplication, and the index $(-i)$ denotes everything else other than i .

Single cell data. As for posterior inference of S, κ, τ in single cell data, note that the first part of the model can be re-written as

$$\begin{aligned} (\kappa_l, \tau_l) &\sim N(\mu, \Sigma), \text{ where } \mu = (\mu_\kappa, \mu_\tau), \Sigma = \text{Diag}(\sigma_\kappa^2, \sigma_\tau^2), \\ S_{il} | \kappa_l, \tau_l &\sim \text{Bernoulli}(\text{logistic}(\psi_{il})), \text{ where } \psi_{il} = \kappa_l + \tau_l A_{i,G_l}, \end{aligned} \quad (3.4)$$

which has the same form as a Bayesian logistic regression, with covariates being $(1, A_{i,G_l})$. Therefore, following the recent development of Gibbs sampling technique in this area (Polson et al., 2013), we introduce a set of augmented latent variables

ω , and the conditional complete posteriors can be shown to be

$$\begin{aligned}\omega_{il} | \omega_{(-il)}, S, Y, \kappa, \tau &\sim \text{Polya-Gamma}(1, \psi_{il}), \\ (\kappa_l, \tau_l) | \kappa_{(-l)}, \tau_{(-l)}, \omega, S, Y &\sim N(m_{\omega l}, V_{\omega l}^{-1}), \\ S_{il} | S_{(-il)}, \omega, S, \kappa, \tau, Y &\sim \text{Bernoulli}(b_{il}),\end{aligned}\tag{3.5}$$

where

$$\begin{aligned}\psi_{il} &= \kappa_l + \tau_l A_{i, G_l}, \\ V_{\omega l} &= \begin{pmatrix} \sum_{i=1}^N \omega_{il} + \sigma_{\kappa}^{-2} & \sum_{i=1}^N \omega_{il} A_{i, G_l} \\ \sum_{i=1}^N \omega_{il} A_{i, G_l} & \sum_{i=1}^N \omega_{il} A_{i, G_l}^2 + \sigma_{\tau}^{-2} \end{pmatrix}, \\ m_{\omega l} &= V_{\omega l}^{-1} \begin{pmatrix} \sum_{i=1}^N S_{il} - N/2 + \mu_{\kappa}/\sigma_{\kappa}^2 \\ \sum_{i=1}^N S_{il} A_{i, G_l} - 1/2 + \mu_{\tau}/\sigma_{\tau}^2 \end{pmatrix}, \\ b_{il} &= \begin{cases} 1, & \text{if } Y_{il} > 0 \\ \text{logit} \left(\psi_{il} + R_l \log \left(\frac{\sum_{n \neq i} A_{n, G_l} S_{nl}}{A_{i, G_l} + \sum_{n \neq i} A_{n, G_l} S_{nl}} \right) \right), & \text{if } Y_{il} = 0 \end{cases}.\end{aligned}$$

3.2. *M-step.* In the M-step of GEM algorithm, the parameters are updated to maximize a lower bound on the expected complete log likelihood function, or the so-called Evidence Lower BOund (ELBO), where the posterior expectation \mathbb{E}_Q is estimated using Gibbs samples obtained in the E-step. The set of optimal dropout parameters are

$$\begin{aligned}\hat{\mu}_{\kappa} &= \frac{1}{L} \sum_{l=1}^L \mathbb{E}_Q(\kappa_l), \quad \hat{\sigma}_{\kappa}^2 = \frac{1}{L} \sum_{l=1}^L \mathbb{E}_Q[(\kappa_l - \hat{\mu}_{\kappa})^2], \\ \hat{\mu}_{\tau} &= \frac{1}{L} \sum_{l=1}^L \mathbb{E}_Q(\tau_l), \quad \hat{\sigma}_{\tau}^2 = \frac{1}{L} \sum_{l=1}^L \mathbb{E}_Q[(\tau_l - \hat{\mu}_{\tau})^2].\end{aligned}\tag{3.6}$$

For A and α , there are no closed form solutions, so the projected gradient ascent algorithm is used:

$$\begin{aligned}A_k^{new} &\leftarrow \text{Proj}(A_k^{old} + t \cdot \nabla ELBO(A_k^{old})), \\ \alpha^{new} &\leftarrow \text{Proj}(\alpha^{old} + t \cdot \nabla ELBO(\alpha^{old})),\end{aligned}\tag{3.7}$$

where the step size t is determined by backtracking line search, and the *Proj* function is the projection onto feasible set where

$$A_{ik} \geq \epsilon_A, \quad \sum_{i=1}^N A_{ik} = 1, \quad \alpha_k \geq \epsilon_{\alpha},\tag{3.8}$$

for some small pre-determined constants $\epsilon_A, \epsilon_\alpha > 0$. The gradients are computed as

$$\begin{aligned} \frac{\partial ELBO}{\partial A_{ik}} &= \sum_{j=1}^M \frac{\mathbb{E}_Q [\tilde{Z}_{ij,k}]}{A_{ik}} + \sum_{l:G_l=k} \left[\frac{Y_{il} \mathbb{E}_Q(S_{il})}{A_{ik}} - \mathbb{E}_Q[\omega_{il} \tau_l^2] A_{ik} - \right. \\ &\quad \left. \frac{\mathbb{E}_Q(S_{il}) R_l}{u_l} + \mathbb{E}_Q \left[\left(S_{il} - \frac{1}{2} \right) \tau_l - \omega_{il} \tau_l \kappa_l \right] \right], \\ \frac{\partial ELBO}{\partial \alpha_k} &= \sum_{j=1}^M \mathbb{E}_Q[\log W_{kj}] + M \left[\Psi \left(\sum_{k=1}^K \alpha_k \right) - \Psi(\alpha_k) \right], \end{aligned} \quad (3.9)$$

where $u_l = \sum_{i=1}^N A_{i,G_l} \mathbb{E}_Q(S_{il})$ and $\Psi(\cdot)$ is the digamma function. More detailed derivations can be found in the appendix.

4. Simulation Results.

4.1. Estimation of profile matrix. In this section, we illustrate that our model achieves an accurate estimation on the profile matrix A by correcting for dropouts and utilizing tissue samples. We simulate a scenario with $N = 200$ genes and $L = 100$ single cells coming from $K = 3$ different cell types, with 20, 20, and 60 cells in each type respectively. The sequencing depths are assumed to be $5N$. Recall that the dropout probability at gene i in cell l is determined by

$$1 - \pi_{il} = 1 - \text{logistic}(\kappa_l + \tau_l A_{i,G_l}), \quad (4.1)$$

and here we generate the variables from $\kappa_l \sim N(-1, 0.1^2)$, $\tau_l \sim N(300, 50^2)$. The simulated dropout curves for the 100 single cells are shown in Figure 2a.

A naive method to estimate the profile matrix A is to use the sample means of single cell expression levels, after normalizing by their sequencing depths. Specifically, recall that $Y \in \mathbb{R}^{N \times L}$ represents the observed expression levels in single cells, $\{G_l\}_{l=1,\dots,L}$ represent the cell types, and $\{R_l\}_{l=1,\dots,L}$ are the sequencing depths. Then an entry A_{ik} can be estimated by

$$\hat{A}_{ik}^{naive} = \frac{1}{\#\{l : G_l = k\}} \sum_{l:G_l=k} \frac{Y_{il}}{R_l}. \quad (4.2)$$

However, due to the presence of dropout events and the dependency between π_{il} and A , this naive sample mean estimation is biased (Figure 2b). On the other hand, by explicitly modeling the occurrence of dropout events and capturing the relationships between dropout probability and expected expression level, our model for single cell data provides substantial improvement in estimating A using only the single cell data (Figure 2c).

Moreover, we further simulate $M = 150$ tissue samples, each containing a mixture of the 3 types of cells, with the mixing proportions generated from a Dirichlet distribution with parameter $\alpha = (1, 2, 4)$. The sequencing depths for tissue samples are assumed to be $50N$, which is ten times larger than single cell data as in many

real data sets. Now, after utilizing both single cell and tissue data, our unified model further improves the estimation of profile matrix A , achieving an correlation of 0.997 between true and estimated values (Figure 2d).

4.2. Deconvolution of tissue samples. Now we further examine the model performance on inferring the mixing proportions W in tissue samples, using the same simulation setting as in Section 4.1

We compare the performance of our model to two widely used deconvolution methods, Digital Sorting Algorithm (DSA) (Zhong et al., 2013) and semi-supervised Nonnegative Matrix Factorization (ssNMF) (Gaujoux & Seoighe, 2012). Both DSA and ssNMF rely heavily on a set of “marker genes” as input to guide the matrix factorization, where a marker gene is only expected to be expressed in one cell type. However, marker genes are rarely known in practice, and a widely adopted procedure is to estimate the list of marker genes from purified cells by selecting those with the most different expression levels across cell types. Here, we mimic this procedure by estimating a list of marker genes from single cell data to guide DSA and ssNMF. Specifically, we adopt the method in Abbas et al. (2009), which calculates a p -value of each gene by comparing its expression level in the highest and second-highest types of cells, then selects the group of genes with the smallest p -values. Figure 6 shows the Frobenius loss of DSA and ssNMF using different sets of estimated marker genes with p -values smaller than $\{10^{-6}, \dots, 10^{-14}\}$. It is clear that these two algorithms are sensitive to the input marker genes.

On the other hand, our unified model avoids the estimation of marker genes by automatically utilizing the information in single cell data and achieves lower Frobenius loss than DSA and ssNMF (Figure 6). For comparison, we also show the performances of DSA and ssNMF when the true marker genes are used as input. With this oracle information, both algorithms achieve similar performances as our unified model.

4.3. Inference of dropout entries in single cell data. Finally, we present the inference on dropout entries in single cell data, again using the same setting as in section 4.1. Here, our goal is to distinguish between dropout entries and structural zeros, as defined in equation (2.1). Note that we only need to make inference for locations where the observed expression levels are zero, i.e., $\{(i, l) : Y_{il} = 0\}$. Recall that $S_{il} = 0$ if gene i is dropped out in cell l , and our model provides the estimated posterior mean of S :

$$\tilde{\pi}_{il} = \mathbb{E}(S_{il} \mid X, Y, \theta), \quad (4.3)$$

where θ contains the model parameters. Hence a natural approach is to predict the entries with small $\tilde{\pi}_{il}$ to be dropouts.

A potential competitor for imputing dropout entries is the Nonnegative Matrix Factorization (NMF) (Lee & Seung, 2001). One can construct a low-rank approximation to the single cell expression matrix $Y \in \mathbb{R}^{N \times L}$ using NMF. Intuitively, the approximated values tend to be higher at dropout entries, and closer to zero at structural-zero entries. Interestingly, as shown in Figure 4a, this simple NMF-based method demonstrates a competitive ability to distinguish between dropout genes

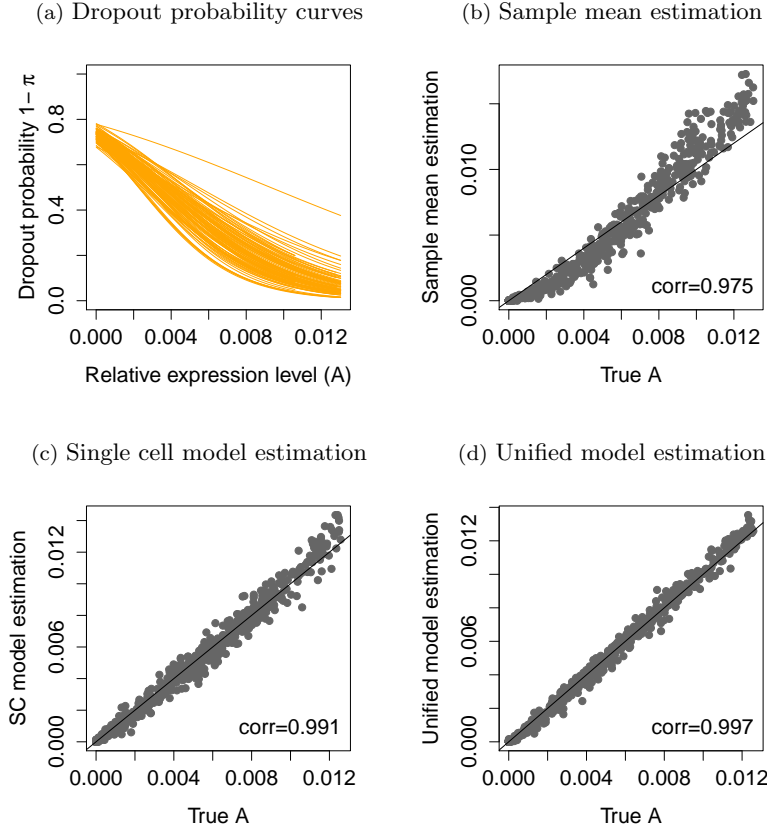


FIG. 2. **(a)** the simulated logistic dropout curves for each single cell, as defined in equation (4.1); **(b)** the true profile matrix A versus the naive sample mean estimation using equation (4.2); **(c)** the true profile matrix A versus the estimation obtained from part of our model using only the single cell data; **(d)** the true profile matrix A versus the estimation obtained from our unified model that also utilizes tissue data.

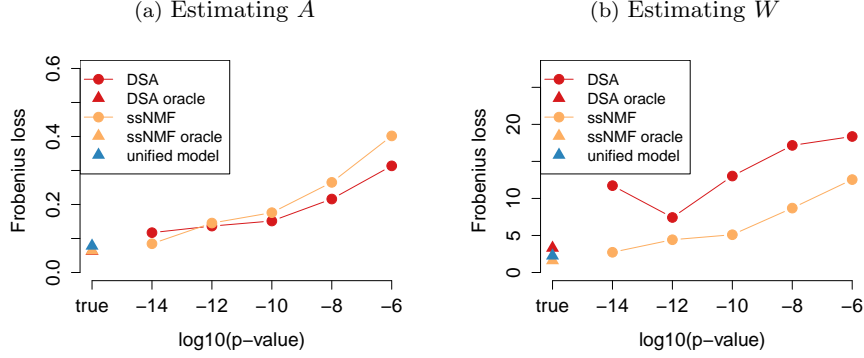


FIG. 3. The Frobenius loss of recovering (a) profile matrix A and (b) mixing proportions W , using DSA and ssNMF, when the marker genes are extracted from single cell data using different thresholds of p -values, as well as under the oracle condition where the true marker genes are given. The performance of our unified model is plotted with a blue triangle in both panels. Our approach does not depend on thresholding p -values.

and structural zeros, if the number of ranks are properly chosen. However, in order to further impute the dropout entries, a good estimation of the profile matrix A is also needed. Figure 4b shows the estimation of A by taking sample average as in equation (4.2), with Y substituted by the NMF approximation. It can be seen that the NMF approach fails to correct for the bias introduced by the dropout events, while our unified model succeeds in both identifying dropout entries and obtaining an unbiased estimation of A (Figure 2d).

5. Application to Fetal Brain Data.

5.1. *Data pre-processing.* In this section, we apply our unified framework to gene expression measured on fetal brains. The single cell RNA-seq data comes from Camp et al. (2015), where 226 cells from fetal brains are sequenced on 18,927 genes. After removing endothelial cells and interneurons, 220 single cells are labeled into three types: 40 apical progenitors (APs), 19 basal progenitors (BPs), and 161 neurons (Ns). The RNA-seq counts are normalized using FPKM (Fragments Per Kilobase of exon per Million fragments mapped) and log-transformed for normality. On the other hand, microarray tissue gene expression data on fetal brains is provided by the BrainSpan atlas (Kang et al., 2011). Within the same window of development, 12 to 13 post-conception week, 72 tissue samples from prefrontal cortex are measured on 16,947 genes. To apply our model, the single cell RNA-seq data are transformed back to linear scale by $2^x - 1$, and all measurements are truncated to integers. To approximate the RNA-seq counts in tissue samples, we transform the BrainSpan microarray data in the same way and treat them as pseudo-RNA-seq counts. The resulting tissue samples have an average pseudo sequencing depth of 5.5×10^6 , which is 26 times larger than the average effective sequencing depth in

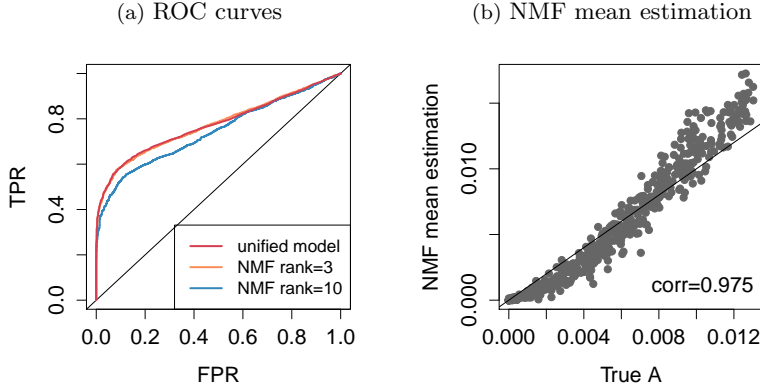


FIG. 4. **(a)** ROC curves of identifying dropout entries in single cell data. **(b)** True profile matrix A versus the sample average of a rank-3 NMF approximation.

single cells, 2.1×10^5 , where the effective sequencing depth is calculated as the sum of FPKM across all genes in each single cell.

To reduce computational load, we only focus on genes with significantly different expression levels among the three cell types. Specifically, we use the 315 so-called PC genes proposed in [Camp et al. \(2015\)](#), which have the largest loadings in a Principal Component Analysis (PCA) and account for the majority of cellular variation. After restricting to the overlapping genes that are also sequenced in BrainSpan tissue samples, a final list of 273 PC genes are obtained and used in the following analyses. When restricting to these 273 genes, the average effective sequencing depth (i.e., the sum of RNA-seq counts in each sample) is 3.2×10^5 ($sd = 1.6 \times 10^4$) in BrainSpan tissues, and 1.4×10^4 ($sd = 4.3 \times 10^3$) in single cells.

Due to the nature of active cell development from APs and BPs to Neurons in fetal brains, we expect to have a few cells that are actively transitioning between two cell types, whose labels are ambiguous. We first remove these ambiguously labeled cells from our analysis. Specifically, we visualize the single cells in the leading 2-dimensional principal subspace obtained by PCA, where the pseudo developing time is constructed using the Monocle algorithm ([Trapnell et al., 2014](#)). Based on the results, the 3 BPs that are close to AP or Neuron clusters are removed, so are the 4 Neurons that are close to AP or BP clusters (Figure 5b). The remaining 213 single cells are retained for analysis, and their gene expression levels on the 273 PC genes are visualized in Figure 5a.

5.2. Imputation of single cell data. Here, we apply our unified framework to identify and impute the dropout entries in single cell data. Note that in order to distinguish between dropout entries and structural zeros in single cell data (equation (2.1)), we only need to focus on the entries where the observed gene expression levels are zero. The inference of dropout entries is based on the estimated posterior expectation of $\mathbb{E}(S_{il} | X, Y, \theta)$. As a result, among the 37,771 zero-observation

entries, 45.7% are inferred to be dropouts with probability one (Figure 5c). These entries are then imputed by their expected values, calculated using the corresponding entries in the estimated profile matrix A multiplied by the sequencing depths of the corresponding cells. To illustrate the impact of imputation, we apply PCA again on the imputed data, and Figure 5d shows that the clusters for different cell types are more clearly separated when visualized using the first two principal components.

5.3. Deconvolution of tissue samples. Finally, we present the deconvolution results of tissue samples using our unified framework. According to the prior knowledge that the proportions in tissue samples should be roughly consistent with that in single cell data, the mixing parameter α is initialized at $(2 \times 10^4, 10^4, 7 \times 10^4)$ for AP, BP and Neurons. The scale of α is chosen to be comparable to the average effective sequencing depths of 1.4×10^4 among all single cells. Figure 6a shows the inferred mixing proportions of APs, BPs and Neurons in each of the 72 tissue samples, with an average of 17.7% AP cells, 8.7% BP cells and 73.6% Neurons.

For comparison, we also apply the Digital Sorting Algorithm (DSA) (Zhong et al., 2013) and semi-supervised Nonnegative Matrix Factorization (ssNMF) (Gaujoux & Seoighe, 2012) on the BrainSpan tissue samples. The marker genes for these two algorithms are selected by comparing each gene’s expression levels in the highest and second-highest types of cells in the single cell data, and genes with p -value $< 10^{-5}$ are treated as markers (Abbas et al., 2009). This procedure leads to 21 AP markers, 6 BP markers and 28 Neuron markers, which serve as input to DSA and ssNMF. Figure 6b and Figure 6c suggest that the proportions estimated by ssNMF tend to have too large variations, while DSA overestimates neural composition, but captures similar patterns for Neuron proportions as our model.

As another perspective to verify the deconvolution results, we use the intuition that the true proportions of a cell type should be correlated with the expression levels of its marker genes in tissue samples. To check whether this holds in the results, we first normalize each tissue sample by their effective sequencing depths, such that the normalized expressions sum to one in each sample. We focus on 7 genes based on biological knowledge, including the radial glia (RG) markers *PAX6* and *GLI3* that are expected to only express in AP and BP cells, the RG marker *HES1* that is mostly expressed in AP cells, the early BP marker *HES6*, as well as neuronal genes *NEUROD6*, *BCL11B* and *MYT1L* (Camp et al., 2015). Table 1 summarizes the correlations calculated by estimated proportions using different methods, and we see that our unified model in general systematically achieves the highest correlations, while ssNMF fails to identify BP cells.

6. Discussion. In this paper, we propose a unified statistical framework to jointly analyze two types of RNA-seq data: the single cell data and the tissue-level data. Our model utilizes the strengths from both data sources, provides a more accurate estimation of cell type specific gene expression profiles, and successfully corrects for the technical noise of dropout events in single cell data. As a side product, our model also achieves deconvolution of tissue data by automatically incorporating the cellular gene expression patterns.

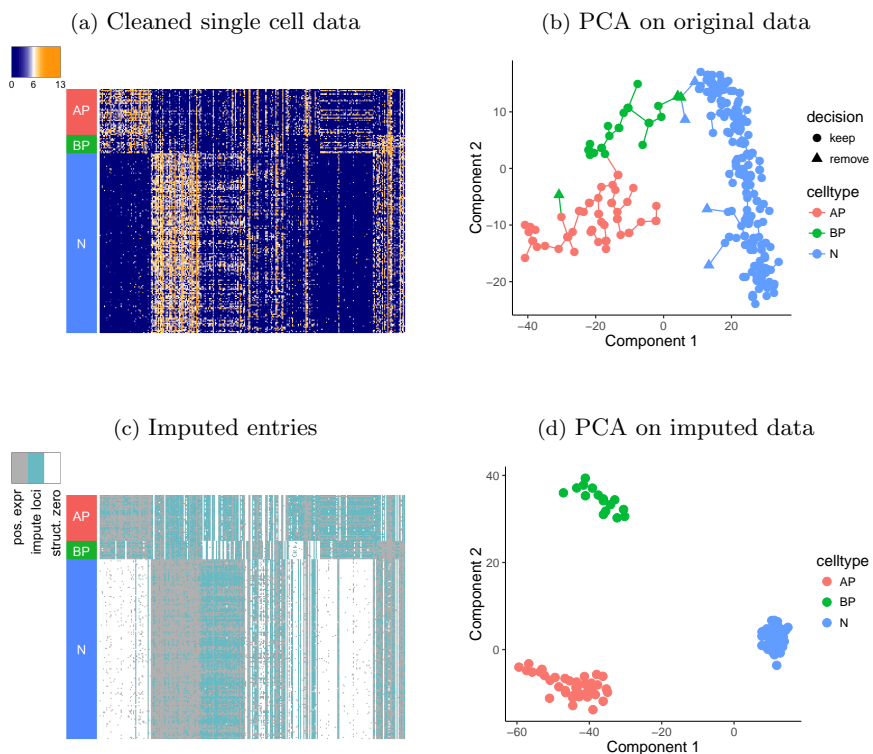


FIG. 5. **(a)** Single cell gene expressions ($\log_2(\text{FPKM}+1)$) after removing 7 ambiguously labeled cells. Rows are 213 cells and columns are 273 genes. **(b)** PCA applied on the original single cell RNA-seq data with 220 labeled cells using 273 PC genes, where the Monocle algorithm is applied to construct pseudo developmental times. 7 cells are identified to be ambiguously labeled and are removed from our analyses (marked as triangles). **(c)** Entries in cleaned single cell data that are inferred to be dropout and imputed (marked in blue) versus the entries that are inferred to be structural zeros (marked in white) in cleaned single cell data. The entries with positive expression levels have no need for posterior inference, and are marked in grey. **(d)** After imputing dropout genes using our unified framework, PCA is conducted on the 213 cells using 273 PC genes, and the resulting cell type clusters are more clearly separated.

TABLE 1

Correlation between the estimated proportions of a cell type in tissue samples and the expression levels of its marker genes. The expression levels have been normalized by sequencing depths. For genes that mark both AP and BP cells, the sum of estimated proportions are used.

Gene	Marked cell type	Unified model	ssNMF	DSA
HES1	AP	0.73	0.68	0.80
HES6	BP	0.66	-0.72	0.53
PAX6	AP.BP	0.91	0.61	0.80
GLI3	AP.BP	0.90	0.54	0.83
NEUROD6	Neuron	0.28	-0.36	0.02
BCL11B	Neuron	0.45	0.02	0.23
MYT1L	Neuron	0.44	0.80	0.32

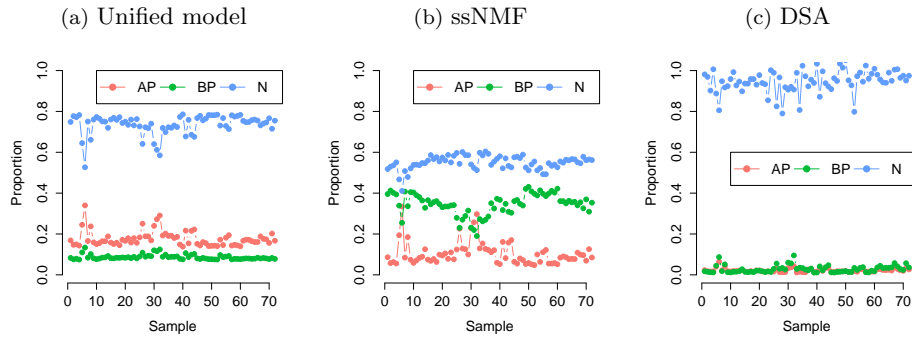


FIG. 6. Deconvolution of tissue samples into three cell types, using (a) our unified framework; (b) semi-supervised Nonnegative Matrix Factorization; (c) Digital Sorting Algorithm.

In the previous sections, we present our model assuming a given number of cell types K . In the situation where K is not known a priori, one can first run the model using a larger value of K , examine the clustering of single cells after imputation, and then reduce to a reasonable choice of K by combining cells from similar clusters.

We apply our framework to two gene expression data sets from fetal brains, and obtain promising results on imputing single cell RNA-seq data and deconvolving tissue samples. With more upcoming single cell data on fetal brains, it would be of great scientific interest to apply our model to specimen from different brain developing periods, which will aid our understanding on gene expression patterns during early brain development and their impact on many complex human disorders.

Acknowledgements. This work was supported by the grants from Simons Foundation (SF402281 and SFARI 124827), National Institute of Mental Health R37MH057881 (Kathryn Roeder and Bernie Devlin), and National Science Foundation (DMS-1553884 and DMS-1407771) (Jing Lei). We thank Bernie Devlin for helpful discussions and comments on the manuscript.

References.

- Abbas, A. R., Wolslegel, K., Seshasayee, D., Modrusan, Z., & Clark, H. F. (2009). Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PLoS One*, 4(7), e6098.
- Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2016). Variational inference: A review for statisticians. *arXiv preprint arXiv:1601.00670*.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993–1022.
- Brennecke, P., Anders, S., Kim, J. K., Kołodziejczyk, A. A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S. A., Marioni, J. C., et al. (2013). Accounting for technical noise in single-cell rna-seq experiments. *Nature methods*, 10(11), 1093–1095.
- Camp, J. G., Badsha, F., Florio, M., Kanton, S., Gerber, T., Wilsch-Bräuninger, M., Lewitus, E., Sykes, A., Hevers, W., Lancaster, M., et al. (2015). Human cerebral organoids recapitulate gene expression programs of fetal neocortex development. *Proceedings of the National Academy of Sciences*, 112(51), 15672–15677.
- Casella, G., & George, E. I. (1992). Explaining the gibbs sampler. *The American Statistician*, 46(3), 167–174.

- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 39(1), 1–38.
- Donoho, D., & Stodden, V. (2003). When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in neural information processing systems*.
- Fan, H. C., Fu, G. K., & Fodor, S. P. (2015). Combinatorial labeling of single cells for gene expression cytometry. *Science*, 347(6222), 1258367.
- Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., Slichter, C. K., Miller, H. W., McElrath, M. J., Pric, M., Linsley, P. S., & Gottardo, R. (2015). Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. *Genome Biol*, 16, 278.
- Gaujoux, R., & Seoighe, C. (2012). Semi-supervised nonnegative matrix factorization for gene expression deconvolution: a case study. *Infection, Genetics and Evolution*, 12(5), 913–921.
- Gelfand, A., & Smith, A. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410), 398–409.
- Geman, S., & Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6(6), 721–741.
- Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1), 5228–5235.
- Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., Clevers, H., & van Oudenaarden, A. (2015). Single-cell messenger rna sequencing reveals rare intestinal cell types. *Nature*, 525(7568), 251–255.
- Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning*, 37(2), 183–233.
- Kang, H. J., Kawasawa, Y. I., Cheng, F., Zhu, Y., Xu, X., Li, M., Sousa, A. M., Pletikos, M., Meyer, K. A., Sedmak, G., et al. (2011). Spatio-temporal transcriptome of the human brain. *Nature*, 478(7370), 483–489.
- Kharchenko, P. V., Silberstein, L., & Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nature methods*, 11(7), 740–742.
- Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C., & Teichmann, S. A. (2015). The technology and biology of single-cell rna sequencing. *Molecular cell*, 58(4), 610–620.
- Lee, D. D., & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, (pp. 556–562).
- Padovan-Merhar, O., & Raj, A. (2013). Using variability in gene expression as a tool for studying gene regulation. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 5(6), 751–759.
- Paisley, J. (2010). Two useful bounds for variational inference. *Technical report*.
- Pierson, E., & Yau, C. (2015). Zifa: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome biology*, 16(1), 1.
- Polson, N. G., Scott, J. G., & Windle, J. (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association*, 108(504), 1339–1349.
- Prabhakaran, S., Azizi, E., & Peer, D. (2016). Dirichlet process mixture model for correcting technical variation in single-cell gene expression data. In *Proceedings of The 33rd International Conference on Machine Learning*, (pp. 1070–1079).
- Repsilber, D., Kern, S., Telaar, A., Walzl, G., Black, G. F., Selbig, J., Parida, S. K., Kaufmann, S. H. E., & Jacobsen, M. (2010). Biomarker discovery in heterogeneous tissue samples -taking the in-silico deconfounding approach. *BMC Bioinformatics*, 11, 27.
- Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., & Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nature biotechnology*, 33(5), 495–502.
- Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., & Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, 32(4), 381–386.
- Vallejos, C. A., Marioni, J. C., & Richardson, S. (2015). Basics: Bayesian analysis of single-cell sequencing data. *PLoS Comput Biol*, 11(6), e1004333.
- Vallejos, C. A., Richardson, S., & Marioni, J. C. (2016). Beyond comparisons of means: understanding changes in gene expression at the single-cell level. *Genome biology*, 17(1), 1.
- Vu, T. N., Wills, Q. F., Kalari, K. R., Niu, N., Wang, L., Rantalainen, M., & Pawitan, Y. (2016).

- Beta-poisson model for single-cell rna-seq data analyses. *Bioinformatics*, 32(14), 2128–35.
- Wainwright, M. J., & Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2), 1–305.
- Wang, W., & Carreira-Perpinán, M. A. (2013). Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application. *arXiv preprint arXiv:1309.1541*.
- Zhong, Y., Wan, Y.-W., Pang, K., Chow, L. M., & Liu, Z. (2013). Digital sorting of complex tissues for cell type-specific gene expression profiles. *BMC bioinformatics*, 14(1), 1.

APPENDIX A: DETAILS OF GIBBS SAMPLING

Recall that the latent variables for tissue data and single cell data are conditionally independent given observed data X, Y and parameters, so we discuss the Gibbs sampling algorithms for the two parts separately.

A.1. Gibbs sampling for tissue data. Recall that the tissue data part of the model is equivalent to the following mixture of multinomials:

$$\begin{aligned}
 W_{\cdot j} &\stackrel{i.i.d.}{\sim} \text{Dirichlet}(\alpha), \quad j = 1, \dots, M, \\
 Z_{rj} &\stackrel{i.i.d.}{\sim} \text{Multinomial}(1, W_{\cdot j}), \quad r = 1, \dots, R_j, \\
 d_{rj} &\stackrel{indep.}{\sim} \text{Multinomial}(1, A_{\cdot Z_{rj}}), \quad r = 1, \dots, R_j, \\
 X_{ij} &= \sum_{r=1}^{R_j} I_{\{d_{rj}=i\}}, \quad i = 1, \dots, N, \quad j = 1, \dots, M,
 \end{aligned} \tag{A.1}$$

where Z and d are represented in scalars, i.e., $\mathbb{P}(Z_{rj} = k) = W_{kj}$, $\mathbb{P}(d_{rj} = i) = A_{i, Z_{rj}}$. We further define

$$\tilde{Z}_{ij,k} := \sum_{r: d_{rj}=i} I_{\{Z_{rj}=k\}}, \quad \tilde{Z}_{ij} := (\tilde{Z}_{ij,k}) \in \mathbb{R}^K, \tag{A.2}$$

then the complete likelihood function for tissue data can be written as below:

$$\begin{aligned}
 L_{\text{tissue}}(W, Z, d, X \mid \alpha, A) &= p(W \mid \alpha) p(Z \mid W) p(d \mid A, Z) p(X \mid d) \\
 &= \prod_{j=1}^M \left\{ \Gamma \left(\sum_{t=1}^K \alpha_t \right) \prod_{k=1}^K \frac{W_{kj}^{(\alpha_k-1)}}{\Gamma(\alpha_k)} \cdot \prod_{r=1}^{R_j} \prod_{k=1}^K W_{kj}^{I_{\{Z_{rj}=k\}}} \right. \\
 &\quad \cdot \prod_{k=1}^K \prod_{i=1}^N A_{ik}^{\sum_{r: d_{rj}=i} I_{\{Z_{rj}=k\}}} \cdot \sum_{i=1}^N I_{\{X_{ij}=\sum_{r=1}^{R_j} I_{\{d_{rj}=i\}}\}} \left. \right\} \\
 &= \prod_{j=1}^M \left\{ \Gamma \left(\sum_{t=1}^K \alpha_t \right) \prod_{k=1}^K \frac{W_{kj}^{(\alpha_k-1)}}{\Gamma(\alpha_k)} \right. \\
 &\quad \cdot \prod_{i=1}^N \prod_{k=1}^K (W_{kj} A_{ik})^{\tilde{Z}_{ij,k}} \cdot \sum_{i=1}^N I_{\{X_{ij}=\sum_{r=1}^{R_j} I_{\{d_{rj}=i\}}\}} \left. \right\}.
 \end{aligned} \tag{A.3}$$

Then it is easy to verify that for any legal realization of d such that

$$X_{ij} = \sum_{r=1}^{R_j} I_{\{d_{rj}=i\}} \quad \text{for } \forall(i, j), \quad (\text{A.4})$$

we have

$$\begin{aligned} p(W_{\cdot j} \mid \tilde{Z}, X, \alpha, A) &\propto \prod_{k=1}^K W_{kj}^{(\alpha_k + \sum_{i=1}^N \tilde{Z}_{ij,k} - 1)}, \\ p(\tilde{Z}_{ij} \mid W, X, \alpha, A) &\propto \left[\prod_{k=1}^K (W_{kj} A_{ik})^{\tilde{Z}_{ij,k}} \right] I_{\{\sum_{k=1}^K \tilde{Z}_{ij,k} = X_{ij}\}}. \end{aligned} \quad (\text{A.5})$$

Therefore, we immediately have

$$\begin{aligned} W_{\cdot j} \mid \tilde{Z}, X &\sim \text{Dirichlet} \left(\alpha + \sum_{i=1}^N \tilde{Z}_{ij} \right), \\ \tilde{Z}_{ij} \mid W, X &\sim \text{Multinomial} \left(X_{ij}, \frac{A_{i\cdot} \odot W_{\cdot j}}{\sum_{k=1}^K A_{ik} W_{kj}} \right), \end{aligned} \quad (\text{A.6})$$

where \odot denotes element-wise multiplication.

A.2. Gibbs sampling for single cell data. Recall that the first part of the single cell model can be rewritten as a Bayesian Logistic regression (equation (3.4)). Then we can utilize the data augmentation trick following Polson et al. (2013). The key is to notice that the logistic function can be written as mixtures of Gaussians with respect to a Polya-Gamma (PG) distribution:

$$\frac{(e^\psi)^a}{(1 + e^\psi)^b} = 2^{-b} e^{c\psi} \int_0^\infty e^{-\omega\psi^2/2} p(\omega) d\omega, \quad \forall \psi \in \mathbb{R}, \quad (\text{A.7})$$

for any constants $a, b > 0$, where $c = a - b/2$ and $\omega \sim PG(0, 1)$. Plugging in this equation, and let

$$\psi_{il} = \kappa_l + \tau_l A_{i,G_l},$$

the complete likelihood for the single cell data can be written as

$$\begin{aligned}
& L_{sc}(Y, \kappa, \tau, S \mid \mu_\kappa, \sigma_\kappa^2, \mu_\tau, \sigma_\tau^2, A) \\
&= p(Y|S, A)p(S|\kappa, \tau)p(\kappa|\mu_\kappa, \sigma_\kappa^2)p(\tau|\mu_\tau, \sigma_\tau^2) \\
&\propto \prod_{l=1}^L \left\{ \frac{R_l!}{\prod_{i=1}^N Y_{il}!} \prod_{i=1}^N \left[\left(\frac{A_{i,G_l}}{\sum_{n=1}^N S_{nl} A_{n,G_l}} \right)^{Y_{il} S_{il}} \delta_0(Y_{il})^{(1-S_{il})} \right] \right. \\
&\quad \cdot \left. \left[\prod_{i=1}^N \frac{(e^{\psi_{il}})^{S_{il}}}{1 + e^{\psi_{il}}} \right] \cdot \frac{1}{\sigma_\kappa \sigma_\tau} \exp \left\{ -\frac{(\kappa_l - \mu_\kappa)^2}{2\sigma_\kappa^2} - \frac{(\tau_l - \mu_\tau)^2}{2\sigma_\tau^2} \right\} \right\} \\
&\propto \prod_{l=1}^L \left\{ \left[\frac{R_l!}{\prod_{i=1}^N Y_{il}!} \frac{\prod_{i=1}^N (A_{i,G_l})^{Y_{il} S_{il}} \delta_0(Y_{il})^{(1-S_{il})}}{\left(\sum_{n=1}^N S_{nl} A_{n,G_l} \right)^{R_l}} \right] \right. \\
&\quad \cdot \left[\prod_{i=1}^N e^{\psi_{il}(S_{il}-1/2)} \int_0^\infty e^{-\omega_{il}\psi_{il}^2/2} p(\omega_{il}) d\omega_{il} \right] \\
&\quad \cdot \left. (\sigma_\kappa^2 \sigma_\tau^2)^{-1/2} \exp \left\{ -\frac{(\kappa_l - \mu_\kappa)^2}{2\sigma_\kappa^2} - \frac{(\tau_l - \mu_\tau)^2}{2\sigma_\tau^2} \right\} \right\} \\
&\propto \int_0^\infty p(Y, \kappa, \tau, S, \omega) p(\omega) d\omega
\end{aligned} \tag{A.8}$$

where $\delta_0(y) = I_{\{y=0\}}$, and $\omega_{il} \sim PG(0, 1)$ independently. Then following the same arguments as in [Polson et al. \(2013\)](#), we get the conditional posterior distribution for $\omega_{il}, \kappa_l, \tau_l$ as follows:

$$\begin{aligned}
& \omega_{il} \mid \omega_{-(il)}, S, Y, \kappa, \tau \sim \text{Polya-Gamma}(1, \psi_{il}), \\
& (\kappa_l, \tau_l) \mid \omega, S, Y \sim N(m_{\omega l}, V_{\omega l}^{-1}),
\end{aligned} \tag{A.9}$$

where

$$\begin{aligned}
V_{\omega l} &= \begin{pmatrix} \sum_{i=1}^N \omega_{il} + \sigma_\kappa^{-2} & \sum_{i=1}^N \omega_{il} A_{i,G_l} \\ \sum_{i=1}^N \omega_{il} A_{i,G_l} & \sum_{i=1}^N \omega_{il} A_{i,G_l}^2 + \sigma_\tau^{-2} \end{pmatrix}, \\
m_{\omega l} &= V_{\omega l}^{-1} \begin{pmatrix} \sum_{i=1}^N S_{il} - N/2 + \mu_\kappa/\sigma_\kappa^2 \\ \sum_{i=1}^N S_{il} A_{i,G_l} - 1/2 + \mu_\tau/\sigma_\tau^2 \end{pmatrix}.
\end{aligned}$$

The only thing left is the conditional posterior for S_{il} . This can be easily obtained by looking at the un-augmented version of likelihood. Note that S_{il} is binary, and we have

$$\frac{P(S_{il} = 1 \mid S_{-(il)}, Y, \psi)}{P(S_{il} = 0 \mid S_{-(il)}, Y, \psi)} = \frac{(A_{i,G_l})^{Y_{il}} e^{\psi_{il}}}{\delta_0(Y_{il})} \cdot \left(\frac{\sum_{n \neq i}^N S_{nl} A_{n,G_l}}{\sum_{n \neq i}^N S_{nl} A_{n,G_l} + A_{i,G_l}} \right)^{R_l}. \tag{A.10}$$

Therefore,

$$S_{il} \mid S_{-(il)}, \omega, \kappa, \tau, Y \sim \text{Bernoulli}(b_{il}),$$

where

$$b_{il} = \begin{cases} 1, & \text{if } Y_{il} > 0 \\ \text{logit} \left(\psi_{il} + R_l \log \left(\frac{\sum_{n \neq i} S_{nl} A_{n, G_l}}{A_{i, G_l} + \sum_{n \neq i} S_{nl} A_{n, G_l}} \right) \right), & \text{if } Y_{il} = 0 \end{cases}.$$

APPENDIX B: DETAILS OF EM ALGORITHM

B.1. M-step in EM algorithm. Here we give more details about the M-step in the Gibbs-EM (GEM) algorithm. By combining equations (A.3) and (A.8), the expectation of complete log likelihood function can be easily derived. However, the term of $\mathbb{E} \left[\log \left(\sum_{n=1}^N S_{nl} A_{n, G_l} \right) \right]$ makes the optimization complicated. We work around this issue by optimizing a *lower bound* of the objective function. The key step is the following lower bound using Jensen's inequality (Paisley, 2010):

$$-\mathbb{E} \left[\log \left(\sum_n X_n \right) \right] \geq -\log u - \frac{\sum_n \mathbb{E}[X_n] - u}{u}, \quad (\text{B.1})$$

where $u = \sum_n \mathbb{E}[X_n]$, for any random variables X_n 's. Using inequality (B.1), we get the following lower bound of the expected complete log likelihood, using the augmented version:

$$\begin{aligned} \mathbb{E}_Q \left[\log p(X, Y, W, \tilde{Z}, \kappa, \tau, S, \omega | \theta) \right] &\geq \text{const.} + \\ &\sum_{j=1}^M \left\{ \log \Gamma \left(\sum_{k=1}^K \alpha_k \right) + \sum_{k=1}^K [(\alpha_k - 1) \mathbb{E}_Q [\log W_{kj}] - \log \Gamma(\alpha_k)] \right. \\ &\quad \left. + \sum_{i=1}^N \sum_{k=1}^K \left[\mathbb{E}_Q \left(\tilde{Z}_{ij, k} \log W_{kj} \right) + \mathbb{E}_Q \left(\tilde{Z}_{ij, k} \right) \log A_{ik} \right] \right\} + \\ &\sum_{l=1}^L \left\{ \sum_{i=1}^N \mathbb{E}_Q(S_{il}) Y_{il} \log(A_{i, G_l}) - R_l \left(\sum_{i=1}^N \frac{\mathbb{E}_Q(S_{il}) A_{i, G_l}}{u_l} + \log u_l \right) \right. \\ &\quad \left. + \sum_{i=1}^N \mathbb{E}_Q \left[\left(S_{il} - \frac{1}{2} \right) (\kappa_l + \tau_l A_{i, G_l}) - \frac{\omega_{il} (\kappa_l + \tau_l A_{i, G_l})^2}{2} \right] \right. \\ &\quad \left. - \frac{1}{2} (\log \sigma_\kappa^2 + \log \sigma_\tau^2) - \frac{\mathbb{E}_Q [(\kappa_l - \mu_\kappa)^2]}{2\sigma_\kappa^2} - \frac{\mathbb{E}_Q [(\tau_l - \mu_\tau)^2]}{2\sigma_\tau^2} \right\}, \end{aligned} \quad (\text{B.2})$$

where $u_l = \sum_{i=1}^N A_{i, G_l} \mathbb{E}_Q(S_{il})$, and \mathbb{E}_Q is the expectation of the posterior distribution, which can be estimated using Gibbs samples. We omit the constants that only involve the data (X, Y) , since the goal here is to optimize over the parameters $\theta = (A, \alpha, \mu_\kappa, \sigma_\kappa^2, \mu_\tau, \sigma_\tau^2)$. In addition, we use the fact that

$$\mathbb{E}_Q[S_{il}] = 1 \quad \text{when } Y_{il} > 0. \quad (\text{B.3})$$

The above lower bound is also referred to as the Evidence Lower BOUND (ELBO). With (B.2), it is straightforward to derive the derivatives mentioned in section 3.2.

The final missing piece is the projection functions in equation (3.7). The projection function for α is straightforward: for any constant $\epsilon_\alpha > 0$,

$$\text{Proj}(\alpha_k) = \max\{\epsilon_\alpha, \alpha_k\}. \quad (\text{B.4})$$

As for the profile matrix A , the projection function is to project onto a subset of simplex

$$\mathcal{S}_\epsilon = \left\{ u = (u_1, \dots, u_N) \in \mathbb{R}^N : \sum_{n=1}^N u_n = 1, u_n \geq \epsilon, \forall n \right\}, \quad (\text{B.5})$$

for some constant $\epsilon > 0$. Wang & Carreira-Perpinán (2013) propose an efficient algorithm for the case when $\epsilon = 0$. The algorithm can be easily generalized to handle a general ϵ , which is specified below:

Projection algorithm for A . For any vector $v \in \mathbb{R}^N$, constant $\epsilon \geq 0$,

1. Sort v into \tilde{v} , such that $\tilde{v}_1 \geq \tilde{v}_2 \geq \dots \geq \tilde{v}_N$.
2. Find $\rho = \max \left\{ 1 \leq j \leq N : \tilde{v}_j + \frac{1}{j}(1 - \sum_{i=1}^j \tilde{v}_i - (N-j)\epsilon) > \epsilon \right\}$.
3. Let $\lambda = \frac{1}{\rho} (1 - \sum_{i=1}^\rho \tilde{v}_i - (N-\rho)\epsilon)$.
4. Let $v_i^* = \max\{\tilde{v}_i + \lambda, \epsilon\}$, then $v^* = (v_i^*) \in \mathbb{R}^N$ is the projection.

B.2. Starting values of EM algorithm. One concern of our inference and estimation procedure is that EM algorithm may be trapped in local optima. Therefore, it is critical to carefully choose the starting values of parameters for EM. For the profile matrix A , a good candidate is the sample means \hat{A}^{naive} in single cell data, as defined in equation (4.2). As shown in Figure 2b, although being biased, \hat{A}^{naive} is usually not too far away from the true profile matrix A . For α , the starting value can be chosen using prior knowledge. Although the exact mixing proportions for each tissue sample is unknown, scientists usually have a good sense of the rough proportions of different cell types in certain tissues. In the case where prior knowledge is unavailable, α can simply be set to $\mathbf{1}^K$, which corresponds to a uniform distribution. In fact, all the simulations in Section 4 use $\alpha_0 = (1, 1, 1)$ as the starting value, and the performances are satisfactory.

The most challenging part is to choose the starting values for μ_κ and μ_τ , which defines the average shape of the logistic curves for dropout probabilities. Although the overall dropout proportion in single cell data is believed to be roughly around 20%, there is usually very limited prior knowledge about the dependency of dropout probabilities on A . Here, we illustrate how the performance of our model changes when using different initial values of μ_κ and μ_τ .

Under the same setting as in Section 4.1, where the true values are $\mu_\kappa = -1$, $\mu_\tau = 300$, the EM algorithm is conducted using different starting values: (i) $\mu_{\kappa 0} = \{-5, -3, -1, 1, 3\}$ when the initial $\mu_{\tau 0}$ is set to be 300; and (ii) $\mu_{\tau 0} = \{100, 200, 300, 400, 500\}$ when the initial $\mu_{\kappa 0}$ is set to be -1 . As shown in Figure B.1, the model remains stable in terms of distinguishing between dropout entries and structural zeros.

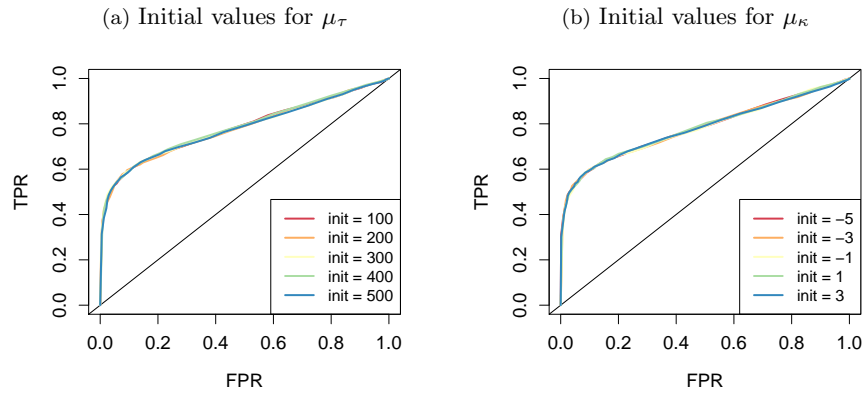


FIG. B.1. ROC curves for identifying dropout entries when starting values of μ_κ and μ_τ change. (a) $\mu_{\tau_0} = \{100, 200, 300, 400, 500\}$ when the initial μ_{κ_0} is set to be -1 ; (b) $\mu_{\kappa_0} = \{-5, -3, -1, 1, 3\}$ when the initial μ_{τ_0} is set to be 300 .

DEPARTMENT OF STATISTICS
 CARNEGIE MELLON UNIVERSITY
 5000 FORBES AVENUE
 PITTSBURGH, PENNSYLVANIA 15213
 USA
 E-MAIL: lzhu@cmu.edu
jinglei@andrew.cmu.edu
roeder@andrew.cmu.edu